# Use of remote sensing data for estimation of winter wheat yield in the United States

L. Salazar [a] , F. Kogan [b] & L. Roytman [c]

[a] Optical Remote Sensing, NOAA CREST Center, City College of
New York, New York

[b] NOAA, National Environmental Satellite Data and Information
Services, Camp Springs, Maryland 20276

[c] City College of New York, Convent Avenue, New York, NY 10031

Available online: 23 Aug 2007

PLEASE SCROLL DOWN FOR ARTICLE

# Use of remote sensing data for estimation of winter wheat yield in the United States

L. SALAZAR*†, F. KOGAN‡ and L. ROYTMAN§

†Optical Remote Sensing, NOAA CREST Center, City College of New York,
New York, USA
‡NOAA, National Environmental Satellite Data and Information Services,
Camp Springs, Maryland 20276, USA
§City College of New York, Convent Avenue, New York, NY 10031, USA

This paper shows the application of remote sensing data for estimating winter wheat yield in Kansas. An algorithm uses the Vegetation Health (VH) Indices (Vegetation Condition Index (VCI) and Temperature Condition Index (TCI)) computed for each week over a period of 23 years (1982–2004) from Advance Very High Resolution Radiometer (AVHRR) data. The weekly indices were correlated with the end of the season winter wheat (WW) yield. A strong correlation was found between winter wheat yield and VCI (characterizing moisture conditions) during the critical period of winter wheat development and productivity that occurs during April to May (weeks 16 to 23). Following the results of correlation analysis, the principal components regression (PCR) method was used to construct a model to predict yield as a function of the VCI computed for this period. The simulated results were compared with official agricultural statistics showing that the errors of the estimates of winter wheat yield are less than 8%. Remote sensing, therefore, is a valuable tool for estimating crop yields well in advance of harvest, and at a low cost.

## 1. Introduction

Recent dry and drought years in the Great Plains have emphasized the need for new sources of timely, objective and quantitative information on crop conditions. Crop growth monitoring and yield estimation can provide important information for government agencies, commodity traders and producers in planning harvest, storage, transportation and marketing activities. The sooner this information is available, the lower the economic risk, translating into greater efficiency and increased return on investments. This paper focuses on wheat because it is by far the world's largest and most widely cultivated food crop. Wheat is the source of 15% to 60% of the calories and protein in the diets of nearly all countries. Bread, the principal product of wheat, is considered the staff of life in most cultures (Shroyer *et al.* 2004).

The major wheat producing countries are China, India, US, France, Russia, Canada and Australia. Marketing of wheat is a multi-billion dollar industry. World demand for wheat is growing 1% per year. Total world wheat production in 2004 reached 630 million tonnes, but in 2003 wheat production was 70 million tonnes less.

---

*Corresponding author. Email: lsalazar@gc.cuny.edu

The major wheat exporting countries, the USA, Canada, France, Australia, Argentina, Germany, United Kingdom and Kazakhstan, supply approximately 70% of the wheat traded in the world market. The major importing countries include Brazil, Egypt, Italy, Japan, Iran, Algeria and China (FAO 2005).

In the USA, wheat is the fourth leading field crop and the leading export crop. Total USA wheat production in 2003 reached 64 million tonnes, while in 2005 wheat production was only 57 million tonnes (FAO 2005). This reduction was due to unfavourable weather. Weather information is normally used to forecast crop yield. However, the weather station network used for these assessments is limited compared to satellite data. For example, the National Oceanic and Atmospheric Administration's (NOAA's) National Weather Service (NWS) oversees 266 weather stations throughout Kansas. Every station provides a spatial coverage of around 800 km$^2$. The purpose of the present research was to use AVHRR data that provides environmental information for every 16 km$^2$.

Tremendous advances in remote sensing technology and computing power over the last few decades are now providing scientists with the opportunity to investigate, measure and model environmental patterns and processes with increasing confidence. Remote sensing of the Earth is playing an increasing role in understanding the natural environment and its inherent physical, biological and chemical processes.

The uses of remote sensing for crop monitoring and yield assessments already represent a very active field of research and application. In Europe, the MARS (Monitoring of Agriculture by Remote Sensing) Project of the Joint Research Centre has taken a leading role in such development (Csornai *et al.* 2002, ITA 2002). In the USA, the United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) uses satellite data to enhance its program of crop acreage estimates. This program is used for construction of the nation's area sampling frame for agricultural statistics, improvement of the statistical precision of crop acreage estimate indicators, especially at the county level and application of GIS based Cropland Data Layer used for watershed monitoring, soil utilization analysis, agribusiness planning, crop rotation practice analysis, animal habitat monitoring and prairie water pothole monitoring (Craig 2001, Mueller *et al.* 2003).

In addition, AVHRR-based vegetation health indices were found to be very useful for early drought detection and for monitoring drought impacts on crop and pasture production around the world, including such major agricultural producers as China, Russia, Brazil, Argentina and Kazakhstan (Dabrowska-Zielinska *et al.* 2002, Kogan 2002, Liu and Kogan 2002, Kogan *et al.* 2003, Domenikiotis *et al.* 2004, Kogan *et al.* 2005). In the USA, these indices were also applied for monitoring corn production in the Great Plains (Hayas and Decker 1996). This paper investigates the application of AVHRR-based vegetation health indices as proxies for the characterization of weather conditions and their impacts on winter wheat yield.

## 2. Study area and data

The study area was Kansas, which is the largest winter wheat (WW) producing state in the US. Nearly one-fifth of all USA WW volume is produced in Kansas (USCRB 2005). Annual average wheat production in Kansas for the past five years has been about 10 million tonnes harvested from an average four million hectares. Kansas ranks number one out of all the US states in wheat and wheat products exported (USCRB 2005). Two types of wheat are grown in the US, winter wheat sown in the autumn and harvested in early summer, and spring wheat planted in the spring and
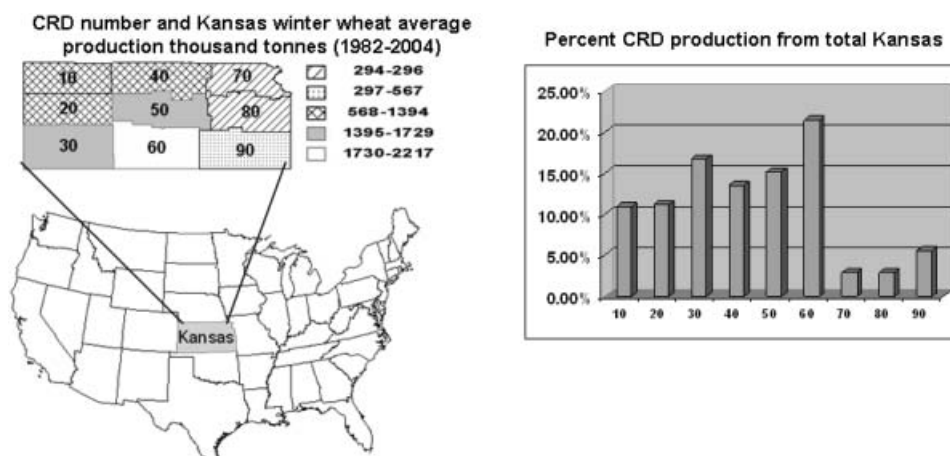
Figure 1.   Area of study, Kansas CRDs and Kansas average winter wheat production (1982–2004) (USCRB 2005).

harvested in late summer/early autumn. Winter wheat provides 70% to 80% of the total wheat production (USCRB 2005). Kansas is divided into nine Crop Reporting Districts (CRDs) as shown in figure 1. From the figure, it can be seen that western and central CRDs are the major producers of WW. CRD 60 is the major producer, followed by CRD 30 and CRD 50.

Kansas has what is typically described as a continental climate, without the influence of any major bodies of water. Annual average precipitation ranges between approximately 102 cm in the southeast to less than 51 cm in the western part of the state, as shown in figure 2. Summers are warm, with most of the annual precipitation occurring during this period.

## 2.1   Winter wheat

Winter wheat production (in tonnes, t), area (in hectares, ha) and yield (in t ha$^{-1}$) were collected from USDA/NASS data for the entire state of Kansas, and for each CRD from 1982 through to 2004 (USDA 2005). Following their methodology, WW production and



Figure 2.   Average annual precipitation in Kansas, 1982 to 2004 (cm) (USHCN 2005).

area were estimated using a sampling technique, and yield was calculated by dividing the total winter wheat production by the area sown (USDA 2005).

## 2.2 Satellite data

Satellite data, including AVHRR-measured solar energy reflected/emitted from the land surface (represented in 8-bit counts), was collected from the NOAA Globa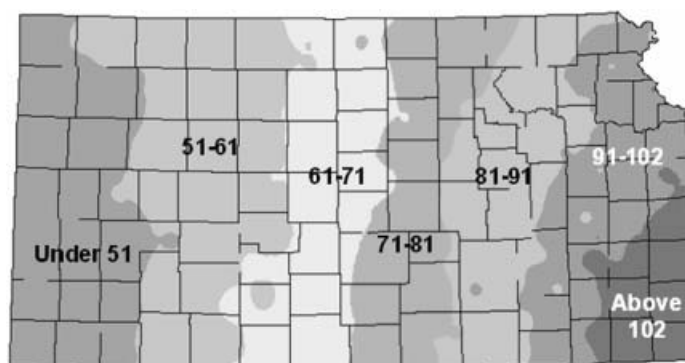l Vegetation Index (GVI) data set from 1982 through to 2004. The GVI data set was developed by sampling $4\,km^2$ Global Area Coverage (GAC) data to $16\,km^2$ spatial resolution and from daily observations to seven-day composite data (Kidwell 1997). The GVI digital counts in the visible (VIS, $0.58$–$0.68\,\mu m$, Ch1), near infrared (NIR, $0.72$–$1.00\,\mu m$, Ch2) and infrared (IR, $10.3$–$11.3\,\mu m$, Ch4) spectral regions were used in this research. Post-launch calibrated VIS and NIR counts were converted to reflectances (Kidwell 1997) and used to calculate the Normalized Difference Vegetation Index ($NDVI = (NIR - VIS)/(NIR + VIS)$). The channel 4 (Ch4) counts were converted to brightness (radiative) temperature (BT) using the method shown in Kidwell (1997).

In order to reduce long-term systematic errors in the GVI time series (Gutman 1999, Kogan and Zhu 2001, Simoniello *et al.* 2004), the following procedure was used. The VIS and NIR channel values were post-launch calibrated following the methods of Rao and Chen (1995, 1996, 1999), Kidwell (1997) and Heidinger *et al.* (2003), and normalized by the cosine solar zenith angle (SZA) and corrected for the Sun–Earth distance. Quality/cloud (QC) masks were developed for each weekly image based on a climatology of channel 4 temperatures (Gutman 1999). For data smoothing a combination of a compound median filter and the least squares technique was applied to the weekly time series. This smoothing completely eliminated high frequency outliers (including random effects), and pulled out low frequency weather related fluctuations (valleys and hills in the NDVI and BT time series) during the annual cycle (Kogan 1997). After smoothing, inter-annual differences due to weather variations in the NDVI and BT data became more apparent (Kogan *et al.* 2003).

The post-launch corrections and time series smoothing considerably improved the stability of the NDVI and BT over time. However, we should admit that there is some remaining long-term noise in the NDVI and BT values. Investigations of this problem by Kogan and Zhu (2001) and Simoniello *et al.* (2004) showed that in most crop related vegetative areas, maximum NDVI and BT change by the end of the satellite life is less than 10%. In addition, this 10% reduction is much smaller than the variation in the NDVI and BT values related to inter- and intra-annual weather changes. This accuracy is appropriate for monitoring vegetation health in vegetative areas (Kogan *et al.* 2003).

Furthermore, previous research showed that when VH indices are correlated with yield anomalies, the correlation coefficient increases considerably during the critical period of crop growth and development. This fact alone indicates that VH indices can be used as proxies for assessment of crop conditions and productivity (Hayas and Decker 1996, Dabrowska-Zielinska *et al.* 2002, Liu and Kogan 2002, Kogan *et al.* 2003, Domenikiotis *et al.* 2004, Kogan *et al.* 2005).

## 3. Methodology

The research strategy of this paper was to extract the weather component from winter wheat yield, NDVI and BT values, and to correlate the weather related component of the yield with the corresponding components of NDVI and BT. The
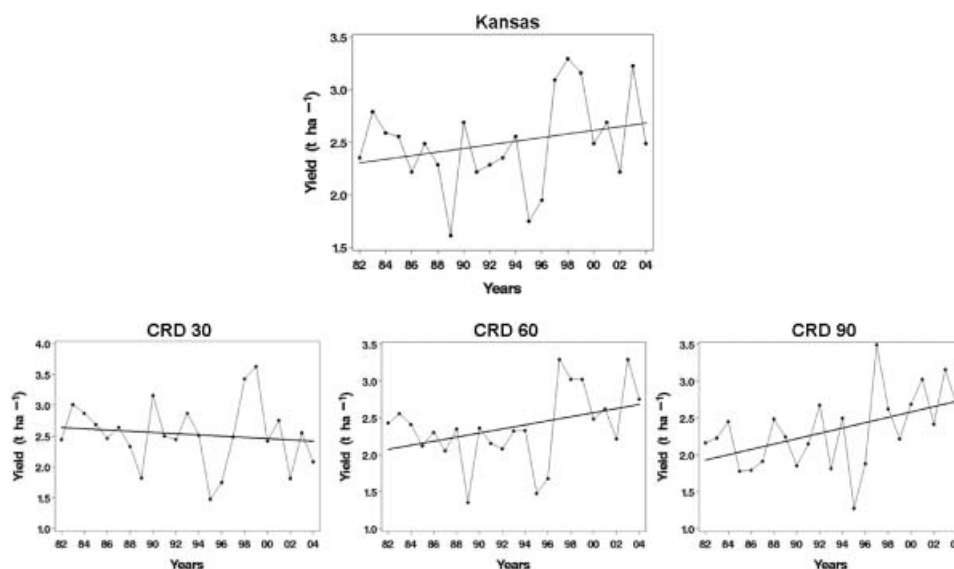
Figure 3.   Winter wheat yield time series.

latter two were expressed in the form of VH indices (Kogan 1997). The goal was to investigate the strength of the relationship and determine if the strongest correlation coincides with the WW's critical period, which is the period when WW production is highly sensitive to weather conditions.

### 3.1   *Winter wheat yield time series*

Following Brockwell and Davis (2000), the WW yield time series shown in figure 3 were approximated by the following equation:

$$Y_t = T_t + \mathrm{d}Y_t, \qquad t = 1, \ldots, \mathrm{n} = 23, \tag{1}$$

where $T_t$ is a slowly changing function representing the deterministic component (trend) that is regulated by agricultural technology, and $\mathrm{d}Y_t$ is a random component regulated by weather fluctuations.

The deterministic component ($T_t$) was estimated using the least squares method. If the yield time series are longer than 30 to 35 years, they might be approximated by a second degree polynomial of the form:

$$T_t = \mathrm{a}_0 + \mathrm{a}_1 t + \mathrm{a}_2 t^2, \tag{2}$$

by choosing the parameters $\mathrm{a}_0$, $\mathrm{a}_1$ and $\mathrm{a}_2$ to minimize $\sum_{t=1}^{n}(Y_t - T_t)^2$. For a shorter time series, as in our case, a linear approximation is sufficient to satisfy the minimum criteria. Figure 3 shows that the linear trend represents tendencies in the WW time series of Kansas CRDs, and the parameters of the linear equations are shown in table 1. The random component ($\mathrm{d}Y_t$) was expressed as a ratio of the observed to the trend estimated yield:

$$\mathrm{d}Y_t = Y_t / T_t. \tag{3}$$

Figure 3 shows that in most areas, WW yield increases; this is due to technology improvement. However, CRD 30 shows a slight decrease in the long-term yield trend.

Table 1. Intercept and slope for winter wheat linear trend yield estimates.

| Region | Kansas | CRD 10 | CRD 20 | CRD 30 | CRD 40 | CRD 50 | CRD 60 | CRD 70 | CRD 80 | CRD 90 |
|---|---|---|---|---|---|---|---|---|---|---|
| Intercept | −477.90 | 757.08 | 482.55 | 330.47 | −1015.92 | −1149.23 | −788.37 | −1945.53 | −1582.48 | −1037.48 |
| Slope | 0.258 | −0.361 | −0.224 | −0.146 | 0.529 | 0.595 | 0.995 | 0.995 | 0.811 | 0.537 |

Although agriculture technology is improving here as well, analysis of the literature indicates that this reduction is related to low precipitation rates in western Kansas (as shown in figure 2) and intensive irrigation practices. Irrigation has stimulated an increase in soil salinity that has become a severe environmental hazard in this region. Farmers are facing decreasing crop yields due, in part, to high levels of salinity. In some areas in western Kansas, land is being taken out of production due to unsustainable crop yields (Miles *et al.* 1977, Hillel 2000, Eldeiry and Garcia 2004).

As can be seen in figure 3, WW yield variations from the trend (d$Y$) in Kansas and the CRDs are large. For example, in Kansas, d$Y$ values in 1989 and 1997 were estimated at 0.65 and 1.24, respectively. This indicated a 35% yield reduction in 1989 due to unfavourable weather, and a 24% increase in 1997 due to favourable weather. These variations might be larger for the CRDs. For example, for the major WW producer in Kansas, CRD 60, d$Y$ in 1989 and 1997 were estimated at 0.57 and 1.38 respectively. This indicated a 43% yield reduction in 1989 due to unfavourable weather, and a 38% increase in 1997 due to favourable weather. In 1989, April and May were the driest on record in many counties in Kansas (see figure 4(*a*)). This contributed to a drought stressed crop. On the other hand, in 1997 spring rainfall was near and above normal (see figure 4(*b*)), which resulted in an above trend WW yield.

### 3.2 *AVHRR-based VH indices*

The principle for constructing VH indices stems from the properties of green vegetation to reflect VIS and NIR, and emit IR solar radiation. If vegetation is healthy, it reflects little radiation in the VIS (due to high chlorophyll absorption of the solar radiation), much in the NIR (due to scattering of light by leaf internal tissues and water content), and emits less thermal radiation in the IR spectral bands (because the transpiring canopy is cooler). As a result, for healthy vegetation, NDVI is large and BT is small. Conversely, for unhealthy vegetation, NDVI is small and BT large (Jensen 2000).

The VH indices were calculated from the NDVI and BT values. Details of the algorithm are presented in Kogan (1997). Here, only important steps are mentioned. These include: (a) the complete elimination of high frequency noise from the NDVI and BT annual time series, (b) the approximation of an annual cycle, (c) the calculation of multi-year climatology, and (d) the estimation of medium to low frequency fluctuations during the seasonal cycle associated with weather variations (i.e. departure from climatology). The Vegetation Condition Index (VCI) characterizing moisture, and the Temperature Condition Index (TCI) characterizing thermal conditions were calculated as:

$$VCI = 100(NDVI - NDVI_{min})/(NDVI_{max} - NDVI_{min}), \tag{4}$$

$$TCI = 100(BT_{max} - BT)/(BT_{max} - BT_{min}), \tag{5}$$
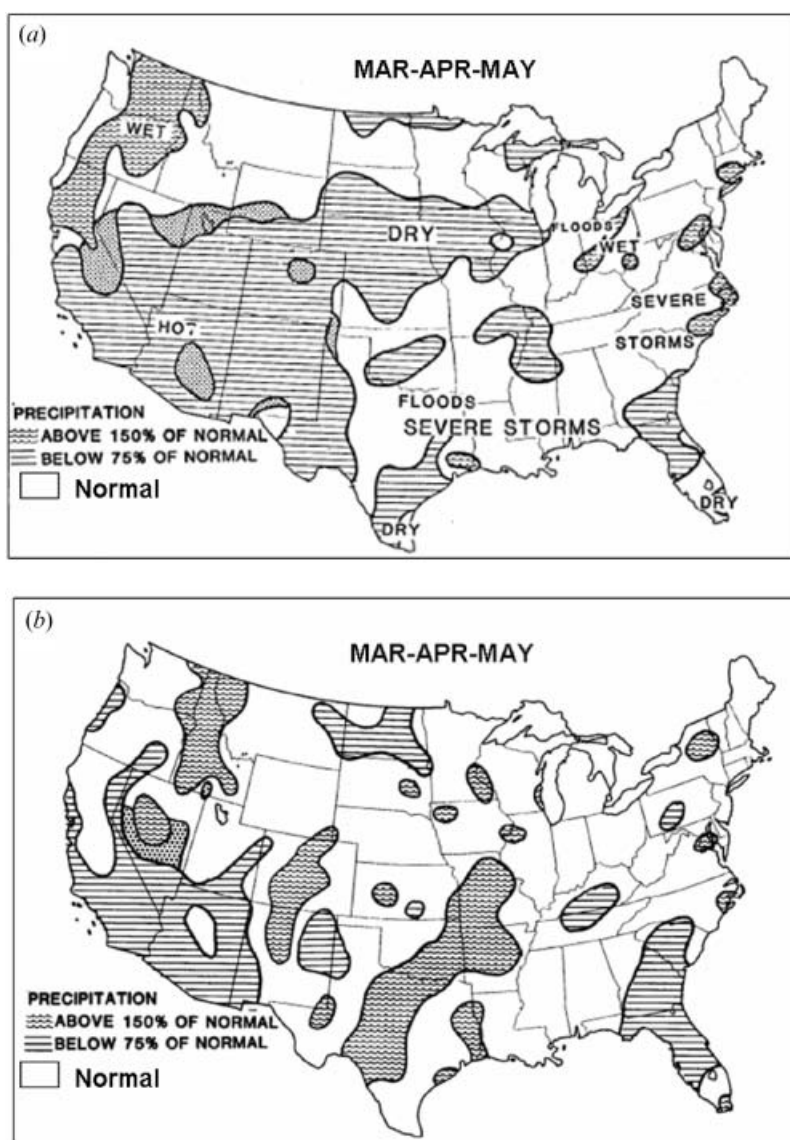
Figure 4. Percentage of normal precipitation in: (*a*) spring 1989, and (*b*) spring 1997 (WWCB 1989, 1997).

where NDVI, $NDVI_{max}$, $NDVI_{min}$, BT, $BT_{max}$ and $BT_{min}$ are the smoothed weekly NDVI or BT values and their 1982 to 2004 absolute maximum and minimum (climatology). The range of VH indices changes from 0, quantifying severe vegetation stress, to 100, quantifying favourable conditions (Kogan 1997). Average weekly values of VH were calculated for each CRD and for total Kansas for the area of WW growth.

## 4. Results and discussion

Since d*Y* and the VH indices were similarly expressed as a deviation from climatology (from trend for yield and from maximum to minimum for VH), further examination
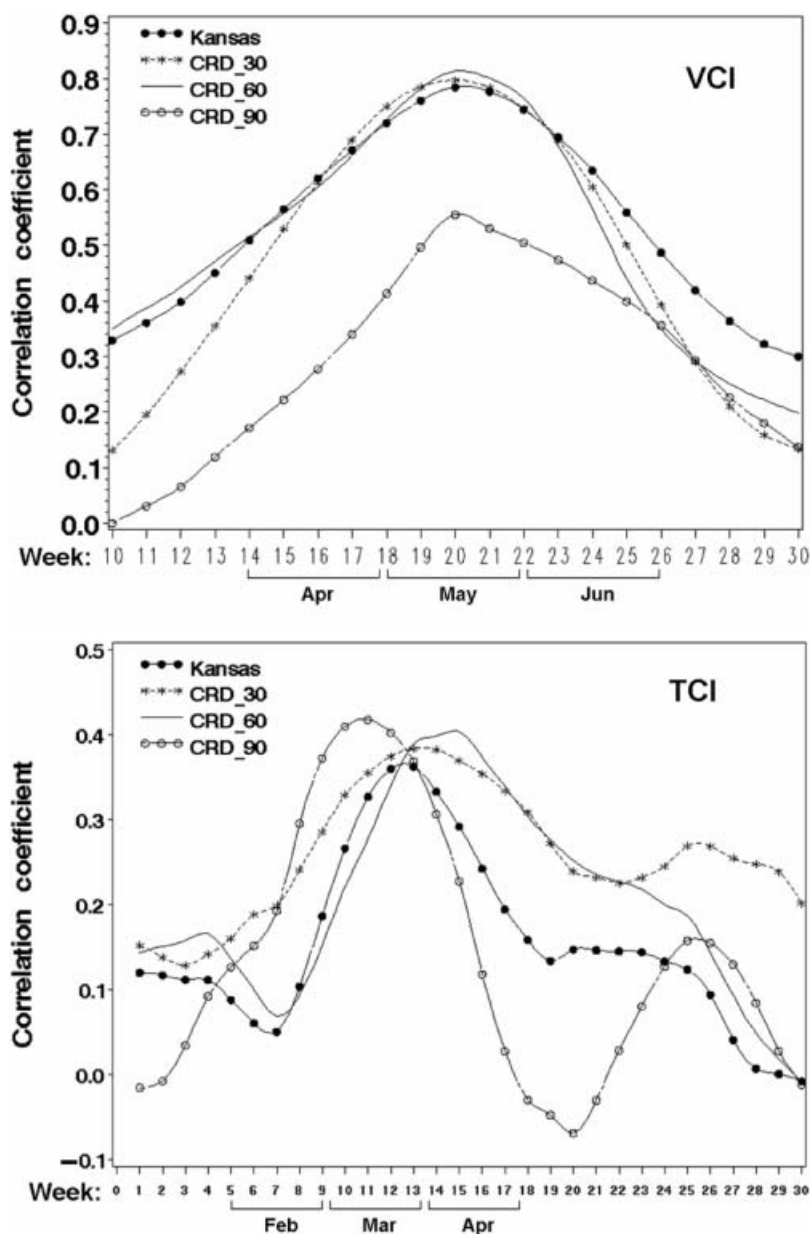
Figure 5. Dynamics of correlation coefficient for d$Y$ versus VCI and TCI.

included correlation and regression analysis of these deviations to investigate the association between them for each CRD and the entire state of Kansas.

Figure 5 shows the dynamics of the correlation coefficients for d$Y$ versus VCI and TCI for Kansas and for three selected CRDs (30, 60 and 90). As seen, d$Y$ is highly correlated with VCI (0.52–0.85) during April, May and early June (weeks 16 to 23). This period is known as critical for yield in Kansas, because WW goes through the reproductive period from the end of the biomass development to the beginning of

maturation. The actual number of kernels that will form in the spike is determined at this stage (Shroyer *et al.* 2004). Positive correlation of d*Y* with a VCI indicates that above trend WW yield is associated with a VCI above 60 (favourable moisture conditions) and below trend yield is associated with a VCI below 40 (moisture stress).

Crop response to moisture and thermal conditions is not equal during the growing season (Kogan *et al.* 2005). According to figure 5, d*Y* dependence on moisture conditions (VCI) is stronger than its dependence on temperature conditions (TCI) during the WW's critical period. However, the highest correlation of d*Y* with TCI is shifted to the earlier time of the growing season (late February and March).

Figure 5 also shows that the d*Y* versus the VCI relationship during the critical period for WW is not equally strong for all regions. For example, CRD 90 that produces three to four times less WW compared to the other two CRDs, has a smaller correlation coefficient (0.52 as opposed to 0.85). This is explained by the fact that the average VCI for CRD 90 was calculated using geographic boundaries that also included those areas where WW is not cultivated.

In the statistical analysis we used both bivariate correlations and multiple regression. The bivariate correlations revealed that d*Y* was significantly related to VCI for weeks 16 to 23 at a $p < 0.05$ significance level. The correlation between d*Y* and TCI, on the other hand, was not significant at a $p < 0.05$ level. Therefore, in multiple regression analysis, d*Y* was regressed on the linear combination of VCI (weeks 16 to 23) values.

The results of fitting the ordinary least squares (OLS) regression model approximated by equation (6) to the state of Kansas and to CRD 60 are shown in table 2.

$$dY = \beta_0 + \beta_1 \times VCI_{16} + \beta_2 \times VCI_{17} + \beta_3 \times VCI_{18} + \beta_4 \times VCI_{19} +$$
$$\beta_5 \times VCI_{20} + \beta_6 \times VCI_{21} + \beta_7 \times VCI_{22} + \beta_8 \times VCI_{23} + \varepsilon \tag{6}$$

Table 2 shows that the value of $R^2$ is large for Kansas (0.86) and CRD 60 (0.92). A comparison of the relative degree of statistical significance of the model with those of the partial regression coefficients reveals multi-collinearity. The overall model is highly significant with *F* values of 12.88 (Kansas) and 12.38 (CRD 60), and *p* values much smaller than 0.001. The smallest *p* value for a partial regression coefficient is 0.0572, which is not significant at a $p < 0.05$ level. This type of result is a natural consequence of multi-collinearity; the overall model may fit the data quite well, but because several independent variables are measuring similar phenomena, it is difficult to determine which of the individual variables significantly contribute to the regression relationship.

The existence of multi-collinearity tends to inflate the variance of predicted values, i.e. predictions of the response variable for sets of independent variables. This inflation may be especially severe when the values of the independent variables are not in the example. In addition, the OLS estimates of the individual regression coefficients tend to be unstable and can affect both inference and model equation (6) forecasting. The estimated values of the coefficients will also be very sensitive to changes in the sample data and to the addition/deletion of a variable in the equation (Chatterjee *et al.* 2000). To avoid this problem, we used an alternative method of estimation, principal components regression (PCR), which results in better estimation and prediction than OLS. This alternative has the potential to produce

Table 2. Results of multiple linear regression (OLS) of d$Y$ on the variables of equation (6).

| Variable | DF | Parameter estimate | Standard error | t value | Pr>\|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 62.54104 | 17.75053 | 3.52 | 0.0078 |
| VCI_W16 | 1 | 6.48299 | 4.70599 | 1.38 | 0.2056 |
| VCI_W17 | 1 | −11.82405 | 14.88928 | −0.79 | 0.4500 |
| VCI_W18 | 1 | 7.73344 | 18.87442 | 0.41 | 0.6928 |
| VCI_W19 | 1 | −1.49892 | 7.80729 | −0.19 | 0.8525 |
| VCI_W20 | 1 | −8.70017 | 7.56735 | −1.15 | 0.2835 |
| VCI_W21 | 1 | 19.05204 | 11.36348 | 1.68 | 0.1321 |
| VCI_W22 | 1 | −14.17956 | 9.37143 | −1.51 | 0.1687 |
| VCI_W23 | 1 | 3.72277 | 3.41033 | 1.09 | 0.3068 |

Kansas: $R^2$=0.86, RMSE=9.52, F=12.88, P<0.001

| Variable | DF | Parameter estimate | Standard error | t value | Pr>\|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 29.30706 | 19.07030 | 1.54 | 0.1629 |
| VCI_W16 | 1 | 3.49046 | 3.86022 | 0.90 | 0.3923 |
| VCI_W17 | 1 | −8.91032 | 10.34932 | −0.86 | 0.4143 |
| VCI_W18 | 1 | 10.63570 | 11.32743 | 0.94 | 0.3752 |
| VCI_W19 | 1 | −6.23895 | 5.08196 | −1.23 | 0.2545 |
| VCI_W20 | 1 | 6.82772 | 4.63947 | 1.47 | 0.1793 |
| VCI_W21 | 1 | −14.96918 | 8.52122 | −1.76 | 0.1170 |
| VCI_W22 | 1 | 16.05870 | 7.34428 | 2.19 | 0.0602 |
| VCI_W23 | 1 | −5.66366 | 2.55124 | −2.22 | 0.0572 |

CRD 60: $R^2$=0.92, RMSE=8.56, F=12.38, P<0.0009

more precision in the estimated coefficients and smaller prediction errors when the predictions are generated using data other than those used for estimation (Draper and Smith 1981, Myers 1986).

Using PCR methodology, the variables in model equation (6) were transformed into a new set of orthogonal or uncorrelated variables called principal components (PCs) of the correlation matrix. This transformation ranks the new orthogonal variables in order of their importance and the procedure then involves eliminating some of the PCs to get a reduction in variance. After elimination of the least important PCs, a multiple regression analysis of the response variable d$Y$ against the reduced set of PCs was performed using OLS estimation. Since the PCs are orthogonal, they are pair-wise independent, and hence OLS is appropriate. Once the regression coefficients for the reduced set of orthogonal variables were calculated, they were mathematically transformed into a new set of coefficients that correspond to the original or initial correlated set of variables in model equation (6). These new coefficients are principal component estimators (Gunst and Mason 1980).

The first part of table 3 shows the eigenvalues of the correlation matrix. From the 'Eigenvalue' column, it is clear that the first principal component has a very large variance (7.23), the second and third have much smaller variances (0.71 and 0.04), and the others have negligible variances. The 'Difference' column gives the differences between adjacent eigenvalues. This statistic shows the rate of decrease in variances of the PCs. The proportion of total variation accounted for by each of the components is obtained by dividing each of the eigenvalues by the total variation.

Table 3. Principal component results for Kansas.

| | Eigenvalues of the Correlation Matrix | | | |
|---|---|---|---|---|
| | Eigenvalue | Difference | Proportion | Cumulative |
| 1 | 7.23297658 | 6.51622157 | 0.9041 | 0.9041 |
| 2 | 0.71675501 | 0.67386858 | 0.0896 | 0.9937 |
| 3 | 0.04288643 | 0.03697766 | 0.0054 | 0.9991 |
| 4 | 0.00590877 | 0.00518869 | 0.0007 | 0.9998 |
| 5 | 0.00072008 | 0.00004161 | 0.0001 | 0.9999 |
| 6 | 0.00067847 | 0.00062417 | 0.0001 | 1.0000 |
| 7 | 0.00005430 | 0.00003393 | 0.0000 | 1.0000 |
| 8 | 0.00002037 | | 0.0000 | 1.0000 |

| | Eigenvectors | | | |
|---|---|---|---|---|
| | Prin1 | Prin2 | Prin3 | Prin4 |
| VCI_W16 | 0.337291 | −0.473839 | 0.602672 | −0.292170 |
| VCI_W17 | 0.351031 | −0.387369 | 0.140501 | 0.105931 |
| VCI_W18 | 0.361843 | −0.264195 | −0.227354 | 0.335778 |
| VCI_W19 | 0.368569 | −0.112244 | −0.402355 | 0.451404 |
| VCI_W20 | 0.370127 | 0.055453 | −0.367292 | −0.372781 |
| VCI_W21 | 0.363545 | 0.238984 | −0.203505 | −0.467141 |
| VCI_W22 | 0.348392 | 0.411122 | 0.105610 | −0.180868 |
| VCI_W23 | 0.325119 | 0.560192 | 0.464753 | 0.443851 |

| | Eigenvectors | | | |
|---|---|---|---|---|
| | Prin5 | Prin6 | Prin7 | Prin8 |
| VCI_W16 | 0.008836 | −0.433733 | −0.042480 | 0.152079 |
| VCI_W17 | −0.067652 | 0.616234 | 0.103690 | −0.548350 |
| VCI_W18 | 0.069462 | 0.290580 | −0.024575 | 0.738220 |
| VCI_W19 | −0.273256 | −0.566044 | 0.009746 | −0.301218 |
| VCI_W20 | 0.650522 | −0.39891 | −0.367107 | −0.162862 |
| VCI_W21 | −0.183231 | 0.012611 | 0.713215 | 0.093158 |
| VCI_W22 | −0.568043 | 0.156306 | −0.560787 | 0.064415 |
| VCI_W23 | 0.369358 | −0.033269 | 0.169765 | −0.033637 |

These quantities are given in the 'Proportion' column. It is obvious that the first component accounts for 90% of the total variation, a result that is typical when a single factor, in this case moisture (VCI), is a common factor in the variability among the original variables. The cumulative proportions printed in the 'Cumulative' column indicate that 99% of the total variation in the eight variables is explained by only two components.

The second part of table 3 ('Eigenvectors') shows the eigenvectors for each of the PCs. These coefficients, which relate the components to the original variables listed on the first column, are scaled so that their sum of squares is unity. This enables which original variables dominate a component to be found. The coefficients of the first PC show a positive relationship with all variables, with somewhat larger contributions from $VCI_{19}$ (0.368) and $VCI_{20}$ (0.370). As expected, these components have the highest correlation coefficient with $dY$ (figure 5) and are in the middle of the critical period of WW. The second component is dominated by $VCI_{23}$ (0.560).

The final yield component, kernel weight, is determined during maturation for WW, which occurs in this week.

The model equation (6) can then be expressed as:

$$Y = \beta_0 + \sum_{i=1}^{8} \beta_i \cdot X_i + \varepsilon. \tag{7}$$

Let $\bar{y}$ and $\bar{x}_j$ be the means of $Y$ and $X_j$, respectively. Also, let $s_y = \left( \sum_{i=1}^{n} (y_i - \bar{y})^2 \big/ (n-1) \right)^{1/2}$ and $s_j = \left( \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \big/ (n-1) \right)^{1/2}$ be the standard deviations of the response and $j$th predictor variable respectively. Equation (7) can then be written in terms of standardized variables as:

$$\widetilde{Y} = \theta_1 \times \widetilde{X}_1 + \theta_2 \times \widetilde{X}_2 + \ldots + \theta_8 \times \widetilde{X}_8 + \varepsilon', \tag{8}$$

where $\widetilde{Y} = (y_j - \bar{y}) / s_y$ is the standardized version of the response variable (d$Y$) and $\widetilde{X} = (x_{ij} - \bar{x}_j) / s_j$ is the standardized version of the $j$th predictor variable (VCI$_j$). The estimated coefficients satisfy:

$$\beta_j = (s_y / s_j) \cdot \theta_j, \quad j = 1, 2, \ldots, 8, \tag{9.1}$$

$$\beta_0 = \bar{y} - \beta_1 \cdot \bar{x}_1 - \beta_2 \cdot \bar{x}_2 - \ldots - \beta_8 \cdot \bar{x}_8. \tag{9.2}$$

The eight principal components of the standardized predictor variables are given by:

$$Z_j = \sum_{i=1}^{8} c_{ij} \cdot X_i, \quad j = 1, \ldots, 8, \tag{10}$$

where $c_{ij}$ are elements of the eigenvectors of the matrix of bivariate correlation between pairs of the explanatory variables. The model in equation (8) may be written in terms of the principal components as:

$$\widetilde{Y} = \alpha_1 \cdot Z_1 + \alpha_2 \cdot Z_2 + \ldots + \alpha_8 \cdot Z_8 + \varepsilon', \tag{11}$$

where the $\alpha$ and $\theta$ values are related by:

$$\alpha_j = \sum_{i=1}^{8} c_{ij} \cdot \theta_i, \quad j = 1, 2, \ldots, 8, \tag{12}$$

or, conversely:

$$\theta_j = \sum_{i=1}^{8} c_{ij} \cdot \alpha_i, \quad j = 1, 2, \ldots, 8. \tag{13}$$

It can be pointed out that just because the first two PCs explain 99% of the variation it does not mean that they form the best subset of predictors for d$Y$ (Hadi and Ling 1998, Jolliffe 2002). Therefore, cross-validation was used to determine the PCs that should be included in the model. For this criterion, the residual of the $i$th observation that results from dropping it and predicting it on the basis of all other observations was computed for each candidate model. The sum of squares of these values is the predicted residual sum of squares, or *PRESS* (Allen 1974, Geisser and

Table 4. Selection of principal components for the prediction based on minimum *PRESS* statistic values.

| Region | VarsInModel | | | | _PRESS_ | _RSQ_ | _ADJRSQ_ | _RMSE_ |
|---|---|---|---|---|---|---|---|---|
| Kansas | Prin1 | Prin4 | Prin6 | Prin7 | 1634.37 | 0.85 | 0.79 | 8.34 |
| CRD 10 | Prin1 | Prin4 | Prin5 | Prin7 | 1097.44 | 0.89 | 0.86 | 7.01 |
| CRD 20 | Prin1 | Prin3 | Prin6 | Prin8 | 3182.98 | 0.86 | 0.81 | 10.57 |
| CRD 30 | Prin1 | Prin3 | Prin6 | Prin8 | 3058.50 | 0.83 | 0.78 | 10.86 |
| CRD 40 | Prin1 | Prin4 | Prin6 | Prin8 | 3298.72 | 0.82 | 0.75 | 12.13 |
| CRD 50 | Prin1 | Prin5 | Prin6 | Prin7 | 2484.77 | 0.87 | 0.83 | 10.69 |
| CRD 60 | Prin1 | Prin3 | Prin4 | Prin7 | 1289.29 | 0.91 | 0.88 | 7.63 |
| CRD 70 | Prin1 | Prin3 | Prin5 | Prin6 | 4493.59 | 0.83 | 0.77 | 14.26 |
| CRD 80 | Prin1 | Prin4 | Prin6 | Prin7 | 4888.22 | 0.77 | 0.68 | 14.07 |
| CRD 90 | Prin1 | Prin4 | Prin5 | Prin8 | 4363.91 | 0.81 | 0.75 | 12.60 |

Eddy 1979),

$$PRESS = \sum_{i=1}^{n} (\hat{e}_i/(1-h_{ii}))^2, \tag{14}$$

where $\hat{e}_i$ and $h_{ii}=x_i(X'X)^{-1}x'_i$ are the residual and the leverage for the *i*th observation in the candidate model, where $x'$ is the transpose of $\chi$ (matrix of VCI predictors) and $x'_i$ is the transpose of column i. Since we have 8 PCs, we proved 255 different models for each region selecting those models with a minimum *PRESS* value as shown in table 4. This table shows that these models explain 85% (Kansas) to 91% (CRD 60) of the variation in d*Y*, almost the same amount as the OLS method explained. It has already been argued that the OLS estimates are unsatisfactory when multi-collinearity is present. Hence, following the PCR analysis, the final set of coefficients for variables in model equation (6) are calculated and presented in table 5.

## 5. Validation of the prediction model (independent testing)

Validation is the step in which the prediction with the chosen model is tested independently. At the beginning of the model building stage, the data were divided into two sets, the training and validation data sets. The division was carried out randomly so that they consisted of 15 and 8 samples respectively. The model selected

Table 5. Estimated regression coefficients for the original variables using models with minimum *PRESS* values.

| Region | Intercept | VCI_W16 | VCI_W17 | VCI_W18 | VCI_W19 | VCI_W20 | VCI_W21 | VCI_W22 | VCI_W23 |
|---|---|---|---|---|---|---|---|---|---|
| Kansas | 76.16 | 3.661 | −2.905 | −3.765 | 3.618 | −7.228 | 18.040 | −14.17 | 3.335 |
| CRD 10 | 95.17 | 11.293 | −22.39 | 9.217 | 1.233 | 5.148 | −9.297 | 10.386 | −5.430 |
| CRD 20 | 66.04 | 10.495 | −24.14 | 15.829 | −13.22 | 44.970 | −65.10 | 44.621 | −12.83 |
| CRD 30 | 73.42 | 1.475 | −12.64 | 25.753 | −32.54 | 32.458 | −16.48 | 2.900 | −0.386 |
| CRD 40 | 29.50 | −4.933 | 20.746 | −28.13 | 8.787 | 0.473 | 21.972 | −26.63 | 8.775 |
| CRD 50 | 62.44 | −5.281 | 24.704 | −35.82 | 15.250 | 9.683 | −10.87 | 5.201 | −2.275 |
| CRD 60 | 20.39 | 1.883 | −4.507 | 5.867 | −5.028 | 9.041 | −16.93 | 16.241 | −5.262 |
| CRD 70 | 42.01 | 1.058 | 0.150 | −1.619 | −2.814 | 9.069 | −2.185 | −7.910 | 5.379 |
| CRD 80 | 80.85 | −6.229 | 19.703 | −17.38 | 0.131 | −1.923 | 20.381 | −20.59 | 6.336 |
| CRD 90 | 46.30 | 16.080 | −35.66 | 16.414 | 4.465 | 10.181 | −19.99 | 9.767 | −0.227 |
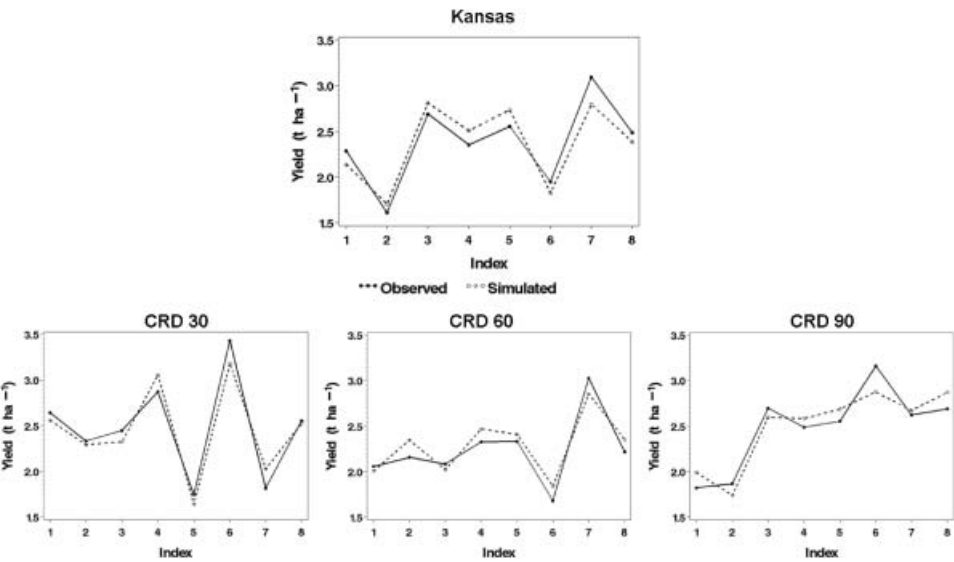
Figure 6. Simulated versus observed winter wheat yield independent testing.

in the optimization step (table 5) was applied to the validation data set and the simulated (S) and observed (O) values were compared. Figure 6 displays observed versus independently simulated WW yield time series. The graphs show that the time series match quite well.

We also regressed the simulated values on the observed values, and statistics were generated. Table 6 shows the statistics of the fit of S values versus O values for WW yield for total Kansas and each CRD. According to Willmott (1982) in 'good' models systematic errors should approach zero, while non-systematic errors should approach the root mean square error (RMSE). Therefore, we can conclude from table 6 that the models based on equation (6) with the coefficients estimated using the PCR methodology detailed in table 5 perform very well.

## 6. Conclusions

In Kansas, the major producer of winter wheat (WW) in the USA, the two AVHRR-based VH indices characterizing moisture (VCI) and thermal (TCI)

Table 6. Statistics of an independent test for the models in table 5.

| Region | Mean absolute error | Root MSE | Systematic error | Non-systematic error | $R^2$ between Simulated (S) and Observed (O) yield |
|---|---|---|---|---|---|
| Kansas | 2.095 | 2.541 | 0.00 | 2.387 | 0.84 |
| CRD 10 | 1.667 | 2.135 | 0.01 | 2.006 | 0.89 |
| CRD 20 | 2.347 | 3.074 | 0.00 | 2.888 | 0.85 |
| CRD 30 | 2.126 | 3.293 | 0.00 | 3.105 | 0.82 |
| CRD 40 | 2.936 | 3.921 | 0.01 | 3.683 | 0.81 |
| CRD 50 | 2.545 | 3.338 | 0.00 | 3.135 | 0.87 |
| CRD 60 | 1.829 | 2.309 | 0.01 | 2.169 | 0.91 |
| CRD 70 | 3.539 | 4.707 | 0.01 | 4.421 | 0.82 |
| CRD 80 | 2.624 | 3.696 | 0.00 | 3.458 | 0.76 |
| CRD 90 | 2.427 | 3.514 | 0.03 | 3.287 | 0.81 |

conditions were tested as predictors of WW yield. It was found that WW was more sensitive to moisture conditions. Correlation analysis between WW yield deviations from trend (d$Y$) with VCI during the period 1982 to 2004, showed strong correlation during the critical period of WW growth (weeks 16 to 23, April to early June). Therefore, this index was used for the statistical modelling of WW yield. This study shows that WW yield can be estimated from the VCI index approximately four weeks prior to harvest time. The VH indices are delivered in real time (every Monday) to http://orbit.nesdis.noaa.gov/smcd/emcb/vci. These results are complementary to crop modelling in other countries (Hayas and Decker 1996, Dabrowska-Zielinska *et al.* 2002, Liu and Kogan 2002, Kogan *et al.* 2003, Domenikiotis *et al.* 2004, Kogan *et al.* 2005). Similar models might be developed for other states, CRDs and even countries. AVHRR data from NOAA polar orbiting satellites can provide valuable information about crop conditions and production on a regional scale in the Great Plains.

Several useful improvements could be made:

(1) To provide conditions and yield estimates for smaller geographic areas. CRD-level yield estimates provide little information about individual counties within each CRD. County-level estimates would improve the ability to identify and assess changes in crop yields for smaller geographic areas. This would improve the overall ability to assess and locate potential production surplus or deficit areas within each CRD. Such information could improve harvest, storage, marketing and crop transportation planning process.

(2) To combine satellite data with weather data specifically during winter and early spring when vegetation is dormant and the application of the VCI is limited.

### References

ALLEN, D.M., 1974, The relationship between variable selection and prediction. *Technometrics*, **16**, pp. 125–127.

BROCKWELL, P.J. and DAVIS, R.A., 2000, *Introduction to Time Series and Forecasting* (New York: Springer).

CHATTERJEE, S., HADI, A.S. and PRICE, B., 2000, *Regression Analysis by Example* (New York: Wiley).

CRAIG, M., 2001, A resource sharing approach to crop identification and estimation. In *Proceedings of the ASPRS 2001 Conference*, Bethesda, MD, USA.

CSORNAI, G., WIRNHARDT, C., SUBS, Z., NADOR, G., MARTINOVITCH, L. and TIKASZ, L., *et al.* 2002, The operational crop monitoring and production forecast program (CROPMON) and other RS based applications. In *Geoinformation for European Wide Integration, Proceedings of the 22nd Symposium of the European Association of Remote Sensing Laboratories*, Prague, Czech Republic (Millpress).

DABROWSKA-ZIELINSKA, K., KOGAN, F., CIOLKOSZ, A., GRUSZCZYNSKA, M. and KOWALIK, W., 2002, Modelling of crop growth conditions and crop yield in Poland using AVHRR-based indices. *International Journal of Remote Sensing*, **23**, pp. 1109–1123.

DOMENIKIOTIS, C., SPILIOTOPOULOS, M., TSIROS, V. and DALEZIOS, N.R., 2004, Early cotton yield assessment by the use of the NOAA/AVHRR derived Vegetation Condition Index (VCI) in Greece. *International Journal of Remote Sensing*, **25**, pp. 2807–2819.

DRAPER, N.R. and SMITH, H., 1981, *Applied Regression Analysis* (New York: Wiley).

ELDEIRY, A. and GARCIA, L., 2004, Spatial modeling using remote sensing, GIS, and field data to assess crop yield and soil salinity. *Hydrology Days*, pp. 55–66.

FOOD AND AGRICULTURE ORGANIZATION (FAO), 2005, Crop production. Available online at: www.fao.org (accessed 14 May 2007).

GEISSER, S. and EDDY, W.F., 1979, A predictive approach to model selection. *Journal of the American Statistical Association*, **74**, pp. 753–760.

GUNST, R.F. and MASON, R.L., 1980, *Regression Analysis and its Application: a Data-Oriented Approach* (New York: M. Dekker).

GUTMAN, G., 1999, On the use of long-term global data of land reflectances and vegetation indices derived from the Advanced Very High Resolution Radiometer. *Journal of Geophysical Research – Atmospheres*, **104**, pp. 6241–6255.

HADI, A.S. and LING, R.F., 1998, Some cautionary notes on the use of principal components regression. *The American Statistician*, **52**, pp. 15–19.

HAYAS, M.J. and DECKER, W.L., 1996, Using NOAA AVHRR data to estimate maize production in the United States Corn Belt. *International Journal of Remote Sensing*, **17**, pp. 3189–3200.

HEIDINGER, A.K., SULLIVAN, J.T. and RAO, C.R.N., 2003, Calibration of visible and near-infrared channels of the NOAA-12 AVHRR using time series of observations over deserts. *International Journal of Remote Sensing*, **24**, pp. 3635–3649.

HILLEL, D., 2000, *Salinity management for sustainable irrigation: integrating science, environment, and economics*, Washington DC, World Bank.

ITA, 2002, *Integrated crop area estimates*, Final MARS/JRC Report.

JOLLIFFE, I.T., 2002, *Principal Component Analysis* (New York: Springer-Verlag).

KIDWELL, K.B., 1997, *Global Vegetation Index User's Guide*, Camp Springs, MD, US Department of Commerce, NOAA, National Environmental Satellite Data and Information Service, National Climatic Data Center, Satellite Data Services Division.

KOGAN, F., 1997, Global drought watch from space. *Bulletin American Meteorological Society*, **78**, pp. 621–636.

KOGAN, F., 2002, World droughts in the new millennium from AVHRR-based Vegetation Health Indices. *Eos*, **83**, pp. 557–564.

KOGAN, F., BANGJIE, Y., GUO, W., PEI, Z. and JIAO, X., 2005, Modelling corn production in China using AVHRR-based vegetation health indices. *International Journal of Remote Sensing*, **26**, pp. 2325–2336.

KOGAN, F., GITELSON, A., ZAKARIN, E., SPIVAK, L. and LEBED, V., 2003, AVHRR-based spectral vegetation indices for quantitative assessment of vegetation state and productivity: calibration and validation. *Photogrammetry Engineering and Remote Sensing*, **69**, pp. 899–906.

KOGAN, F. and ZHU, X., 2001, Evolution of long-term errors in NDVI time series: 1985–1999. *Advances in Space Research*, **28**, pp. 149–153.

LIU, W.T. and KOGAN, F., 2002, Monitoring Brazilian soybean production using NOAA/AVHRR based vegetation condition indices. *International Journal of Remote Sensing*, **23**, pp. 1161–1179.

MILES, D.L., COLORADO STATE UNIVERSITY, COOPERATIVE EXTENSION SERVICE AND US ENVIRONMENTAL PROTECTION AGENCY, REGION VIII, 1977, *Salinity in the Arkansas Valley of Colorado*, Denver.

MUELLER, R., BORYAN, C., CRAIG, M., FLEMING, M. and HANUSCHAK, G., 2003, Pilot Research Project: *Investigation of Very High Resolution Spaceborne Imagery for Citrus Tree Counting*. Report for FDOC Contract No. 02-17.

MYERS, R.H., 1986, *Classical and Modern Regression with Applications* (Boston, Mass.: Duxbury Press).

RAO, C.R.N. and CHEN, J., 1995, Inter-satellite calibration linkages for the visible and near-infrared channels of the Advanced Very High Resolution Radiometer on the NOAA-7, -9 and -11 spacecraft. *International Journal of Remote Sensing*, **16**, pp. 1931–1942.

RAO, C.R.N. and CHEN, J., 1996, Post-launch calibration of the visible and nearinfrared channels of the Advanced Very High Resolution Radiometer on the NOAA-14 spacecraft. *International Journal of Remote Sensing*, **17**, pp. 2743–2747.

RAO, C.R.N. and CHEN, J., 1999, Revised post-launch calibration of the visible and near-infrared channels of the Advanced Very High Resolution Radiometer on the NOAA-14 spacecraft. *International Journal of Remote Sensing*, **20**, pp. 3485–3491.

SHROYER, J.P., WHITNEY, D. and PATERSON, D., 2004, *Wheat production handbook*, Manhattan, Kansas, K-State research and Extension.

SIMONIELLO, T., CUOMO, V., LANFREDI, M., LASAPONARA, R. and MACCHIATO, M., 2004, On the relevance of accurate correction and validation procedures in the analysis of AVHRR-NDVI time series for long-term monitoring. *Journal of Geophysical Research*, **109**, D20107.

US CROP REPORTING BOARD (USCRB), 2005, *Crop production, Washington DC*, Crop Reporting Board Statistical Reporting Service, US Department of Agriculture.

US DEPARTMENT OF AGRICULTURE (USDA), 2005, *Wheat situation and outlook yearbook*, Available online at: www.ers.usda.gov/publications (Washington, DC: USDA).

US HISTORICAL CLIMATOLOGY NETWORK (USHCN), 2005, *Serial Temperature and Precipitation Data*, Environmental Science Division, Publication No. 3404, Carbon Dioxide Information and Analysis Center, Oak Ridge National Laboratory, Oak Ridge, TN, pp. 389.

WEEKLY WEATHER AND CROP BULLETIN (WWCB), 1989, NOAA/USDA, Washington DC, **76**(24) 20 June 1989.

WEEKLY WEATHER AND CROP BULLETIN (WWCB), 1997, NOAA/USDA, Washington DC, **84**(25) 24 June 1997.

WILLMOTT, C.J., 1982, Some Comments on the Evaluation of Model Performance. *Bulletin American Meteorological Society*, **63**, pp. 1309–1313.